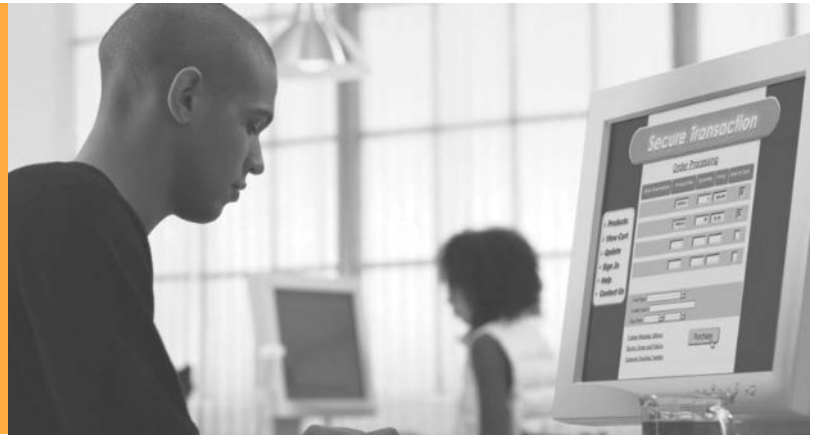
Estadística y Probabilidades



TEMA 2. ANÁLISIS DESCRIPTIVO DE DATOS CON UNA VARIABLE

2.1. Distribuciones de Frecuencias
2.2. Medidas Estadísticas
2.3. Análisis exploratorio de datos con ayuda de herramientas tecnológicas

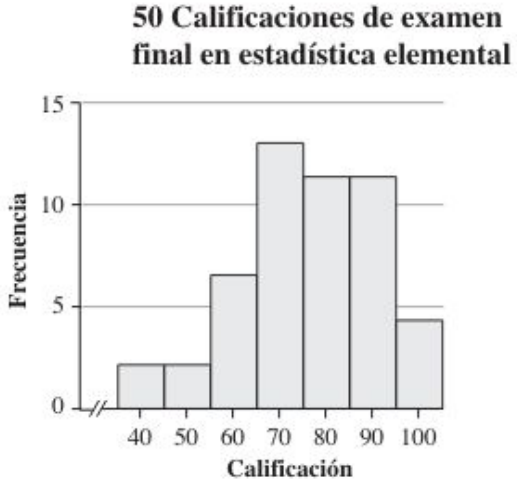
TEMA 2. ANÁLISIS DESCRIPTIVO DE DATOS CON UNA VARIABLE

2.1. Distribuciones de Frecuencias

- Organización de los datos
- Tablas de distribución de frecuencia
- Gráficas

FIGURA 2.1 Gráfica circular

Histograma de frecuencia



Distribución: es el patrón de variabilidad que presentan los datos de una variable. La distribución exhibe la frecuencia de cada valor de la variable.

2.1. Distribuciones de Frecuencias

Organización de los datos

- Ordenamiento
- Frecuencia del dato
- Agrupación de datos en clases
- Tablas de conteo
- Diagrama de tallo y hoja

Tabla estándar para distribución de frecuencia

Número de grupo	Sumas de grupo	Fronteras	Frecuencia
1	||	$35 \leq x < 45$	2
2	||	$45 \leq x < 55$	2
3	||||| ||	$55 \leq x < 65$	7
4	||||| ||||| |||	$65 \leq x < 75$	13
5	||||| ||||| |	$75 \leq x < 85$	11
6	||||| ||||| |	$85 \leq x < 95$	11
7	||||	$95 \leq x \leq 105$	4
			50

Diagrama de tallo y hoja

Calificaciones de 19 exámenes

5	2
6	2 6 8
7	2 4 4 6 6 6 8 8
8	2 2 4 6 8
9	2 6

Diagrama de tallo y hoja: presenta los datos de una muestra usando los dígitos reales que conforman el dato. Cada valor numérico se divide en dos partes: El dígito o dígitos principales son el tallo, y los dígitos que le siguen son las hojas. Los tallos están situados a lo largo del eje principal, y una hoja por cada dato está dispuesta de modo que muestra la distribución de los datos.

2.1. Distribuciones de Frecuencias

Tablas de distribución de frecuencias

- con datos no agrupados y
- con datos agrupados.

Calificaciones de exámenes de estadística

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

Distribución de frecuencia no agrupada

x	*f*
0	1
1	3
2	8
3	5
4	3

Clases	Conteo	x	f	F	r	R
39 - 49	|||	44	3	3	0,06	0,06
49 - 59	||||	54	4	7	0,08	0,14
59 - 69	||||| |||	64	8	15	0,16	0,30
69 - 79	||||| ||||| ||||| |	74	16	31	0,32	0,62
79 - 89	||||| ||||	84	9	40	0,18	0,80
89 - 99	||||| |||||	94	10	50	0,20	1

2.1. Distribuciones de Frecuencias

Procedimiento para construir una distribución de frecuencias agrupadas:

1. Calcular el Rango: Valor Máximo - Valor Mínimo ⇒ Rango = 98 - 39 = 59
2. Estimar el número de grupos o clases: $2^c \geq n$ ⇒ $2^6 \geq 50$ ⇒ c = 6
3. Calcular la amplitud: i_c = Rango / c ⇒ i_c = 59 / 6 = 9.833 ≈ 10
4. Calcular los límites de clase, comenzando con el Valor Mínimo e ir sumando la amplitud:
   a. (Clase 1) Li = Valor Mínimo = 39 Ls = Li + i_c ⇒ Ls = 39 + 10 = 49
   b. (Clase 2) Li = Ls (de la clase anterior), por tanto Li = 49 y Ls = 49 + 10 = 59
   c. (Clase 3) Li = 59 y Ls = 59 + 10 = 69
   d. (Clase 4) Li = 69 y Ls = 69 + 10 = 79
   e. (Clase 5) Li = 79 y Ls = 79 + 10 = 89
   f. (Clase 6) Li = 89 y Ls = 89 + 10 = 99

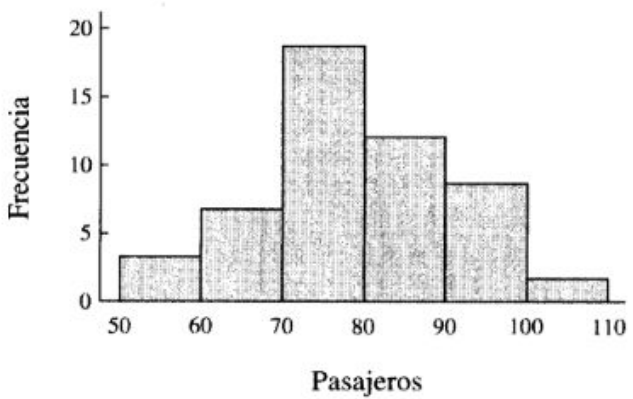
Una vez establecidos los grupos o intervalos de las clases, se cuentan los valores de la tabla de datos que están presentes en cada intervalo y se suman para obtener así la frecuencia absoluta para cada una de las 6 clases en este caso.

Observaciones:

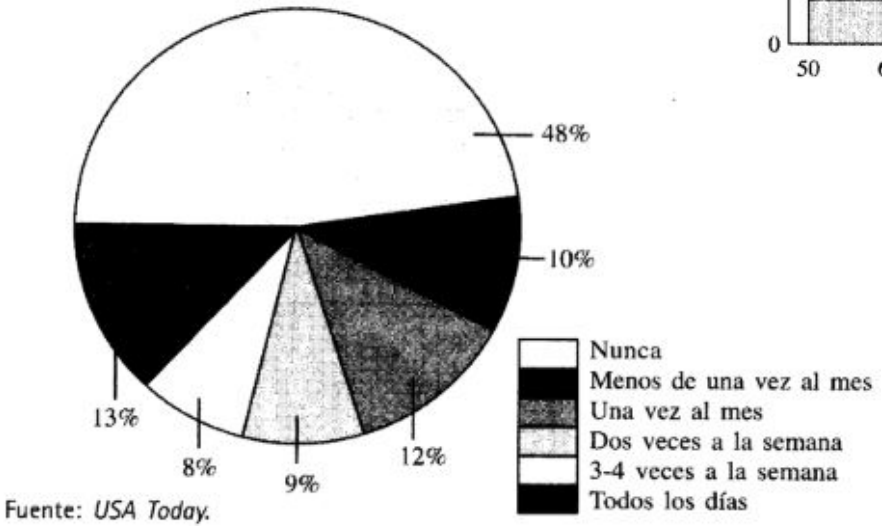
- El Ls de la última clase debe ser mayor o igual al valor máximo.
- Si algún valor de los datos coincide exactamente con el Ls de alguna clases, entonces el valor se suma en la clase siguiente (a excepción de la última clase que si lo debe contar)

2.1. Distribuciones de Frecuencias

Gráficas: Curvas de frecuencias, Histogramas de frecuencias, Polígonos de frecuencias, Ojivas, Diagrama de caja, Diagrama de Pareto.



El Histograma: coloca las clases de una distribución de frecuencias en el eje horizontal y las frecuencias en el eje vertical.

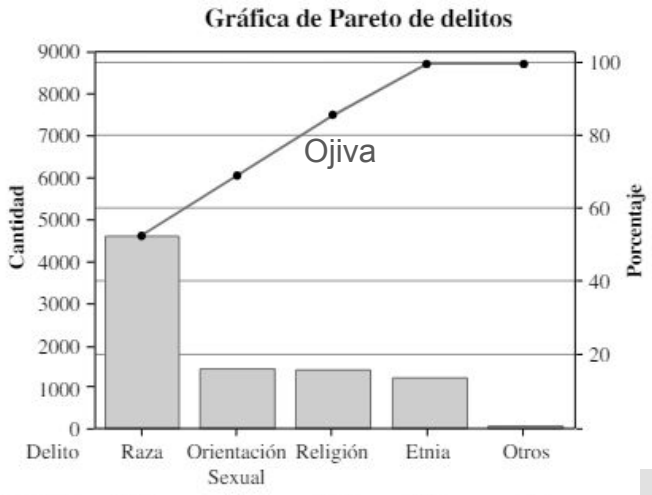


El Diagrama Circular es de especial utilidad para mostrar proporciones relativas (porcentajes) de una variable.

Distribuciones de Frecuencias

Diagrama de Pareto: es una gráfica especial de barras. En ésta, las barras se presentan de la categoría más numerosa a la menos numerosa. Incluye una gráfica de líneas que muestra los porcentajes acumulativos y las cantidades para las barras.

FIGURA 2.3 Diagrama de Pareto

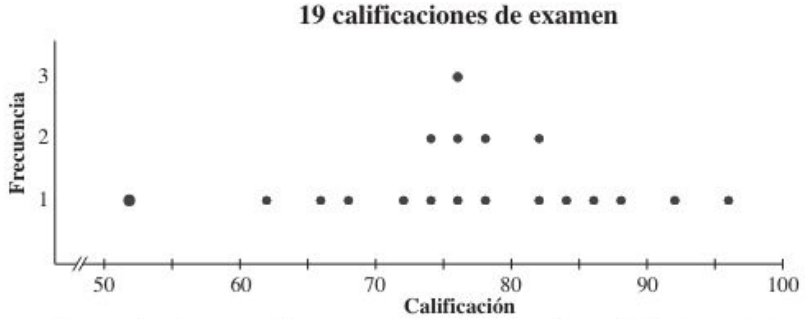


Delito	Raza	Orientación Sexual	Religión	Etnia	Otros
Cantidad	4574	1430	1426	1236	49
Porcentaje	52.5	16.4	16.4	14.2	0.6
Cum %	52.5	68.9	85.3	99.4	100.0

Muestra de 19 calificaciones de examen

76	74	82	96	66	76	78	72	52	68
86	84	62	76	78	92	82	74	88	

La figura 2.4 es una gráfica de puntos de las 19 calificaciones de examen.



Gráfica de puntos: presenta los datos de una muestra al representar cada dato con un punto ubicado a lo largo de una escala que puede ser horizontal o vertical. La frecuencia de los valores se representa a lo largo de la otra escala.

TEMA 2. ANÁLISIS DESCRIPTIVO DE DATOS CON UNA VARIABLE

2.2. Medidas Estadísticas

Medidas de tendencia central:

- media,
- mediana,
- moda,
- otros promedios particulares
  - media ponderada,
  - media geométrica y
  - media armónica.

Medidas de posición u orden:

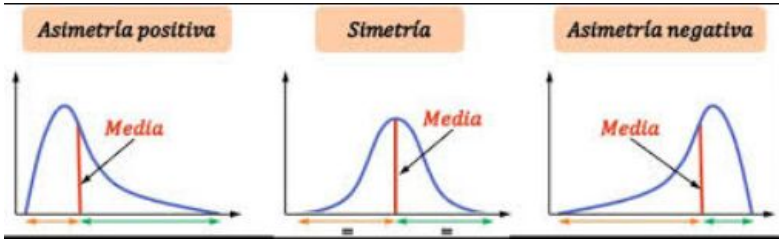
- mediana,
- cuartiles,
- quintiles,
- deciles,
- percentiles.

Medidas de dispersión:

- no referidas a la media aritmética
  - rango o alcance,
  - rangos modificados
- y referidas a la media
  - desviación media,
  - varianza y
  - desviación estándar.

Medidas de forma:

- coeficiente de asimetría,
- coeficiente de curtosis.



Medidas de tendencia central

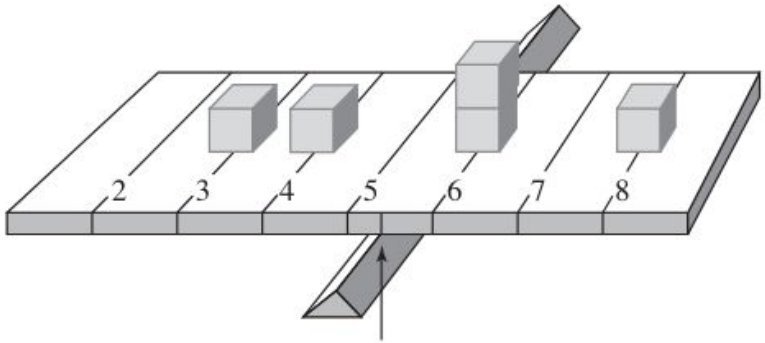
2.2. Medidas Estadísticas

$$\text{media muestral: } \bar{x} = \frac{\sum x}{n}$$

Mediana: es el valor del dato que ocupa la posición media cuando los datos están clasificados en orden de acuerdo con su tamaño.

Un conjunto de datos consta de los cinco valores 6, 3, 8, 6 y 4. Encuentre la media.

$$\bar{x} = \frac{\sum x}{n} = \frac{6 + 3 + 8 + 6 + 4}{5} = \frac{27}{5} = 5.4$$



$\bar{x} = 5.4$ (el centro de gravedad, o punto de equilibrio)

Para **datos no agrupados**, éstos se deben ordenar primero de menor a mayor o viceversa.

Si el número de datos n es impar, la mediana será el valor del dato ubicado en la posición:

$$(n + 1) / 2$$

Si n es par, la mediana será el promedio de los valores de los datos que se encuentran en las posiciones:

$$n / 2 \quad \text{y} \quad (n / 2) + 1$$

2.2. Medidas Estadísticas

Medidas de tendencia central

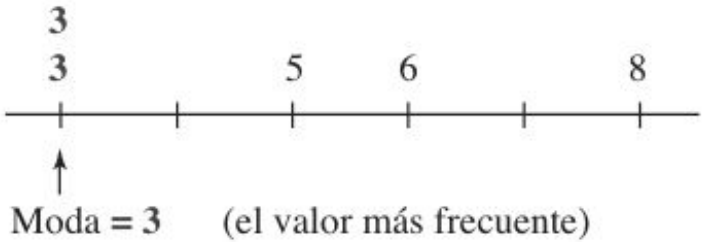
Ejercicio:

Encuentre la mediana para el conjunto:

a) 6, 3, 8, 5, 3
b) 9, 6, 7, 9, 10, 8

La moda: Es el valor de x que se presenta con la mayor frecuencia.

En el conjunto de datos { 3, 3, 5, 6, 8 }
la moda es 3.



Media ponderada: En algunos casos los valores difieren en su grado de importancia o “peso” (w), y se requiere tomar en cuenta estas diferencias al calcular su promedio.

Entonces la fórmula a emplear es así:

$$\text{media ponderada: } \bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$

Ejercicio:

Halle el promedio de tres calificaciones: 8.5, 9.0 y 7.5 con ponderaciones del 20%, 30% y 50% respectivamente.

2.2. Medidas Estadísticas

Medidas de tendencia central

La **media geométrica** suele utilizarse en negocios y economía para calcular las tasas de cambio promedio, las tasas de crecimiento promedio o tasas promedio.

Dados n valores (todos positivos), la media geométrica es la n-ésima raíz de su producto.

Ejercicio:

El factor de crecimiento promedio de dinero compuesto con tasas de interés anual del 10%, el 8%, el 9%, el 12% y el 7% se obtiene determinando la media geométrica de 1.10, 1.08, 1.09, 1.12 y 1.07. Calcule el factor de crecimiento promedio.

$$x = \sqrt[5]{(1.10)(1.08)(1.09)(1.12)(1.07)} = 1.092$$

La **media armónica** se utiliza a menudo como una medida de tendencia central para conjuntos de datos que consisten en tasas de cambios, como la velocidad.

Para calcularla, se divide el número de valores n entre la suma de los recíprocos de todos los valores, de la siguiente forma:

$$\frac{n}{\sum \frac{1}{x}} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}$$

Ejercicio: La media armónica se utiliza para promediar velocidades. Halle la velocidad promedio de un viaje que se realizó a 40 Km/h de ida y 60 Km/h de regreso.

2.2. Medidas Estadísticas

Medidas de posición u orden

Las **medidas de posición** se utilizan para describir la posición que un dato específico posee en relación con el resto de los datos cuando están en orden.

La **mediana** es una medida de posición ya que esta representa el dato que se ubica justo a la mitad de un conjunto de datos ordenados.

Los **cuartiles** son valores de la variable que dividen los datos ordenados en cuartos. Así tenemos:

El Cuartil Q1, marca el primer cuarto o 25%,

El Cuartil Q2, indica el segundo cuarto o 50% y

El Cuartil Q3, tercer cuarto o 75% de los datos.

Los **Quintiles**, similarmente a los cuartiles, representan valores que separan el conjunto de datos ordenados en cinco partes del 20% cada una:

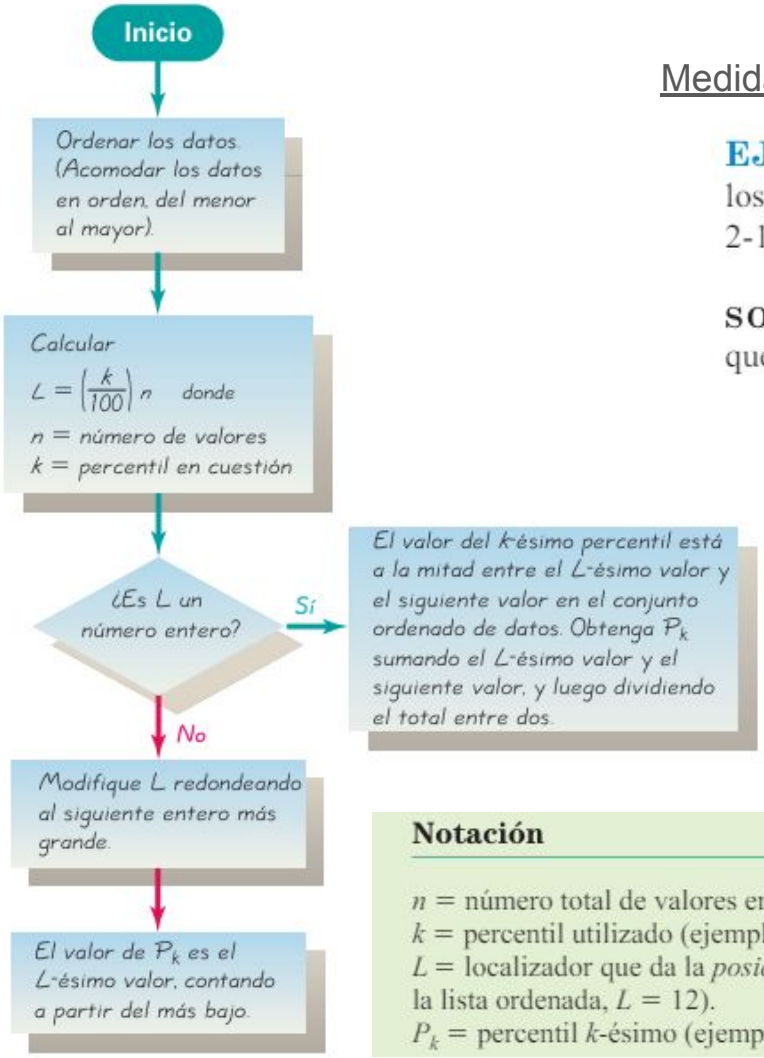
Q1 para el 20%, Q2 para el 40%,

Q3 para el 60% y Q4 para el 80%)

En el caso de los **Deciles**, se tienen 9 valores (D1, D2, …, D9) que separan el grupo de datos en fracciones del 10% cada una.

Y los **Percentiles** 99 valores, desde P1 hasta P99. Cada uno marca un 1% del total.

2.2. Medidas Estadísticas



EJEMPLO Niveles de cotinina de fumadores La tabla 2-13 lista los 40 niveles ordenados de cotinina de fumadores que se incluyen en la tabla 2-1. Calcule el percentil correspondiente al nivel de cotinina de 112.

SOLUCIÓN A partir de la tabla 2-13 se ve que hay dos valores menores que 112; por lo tanto,

$$\text{percentil de } 112 = \frac{12}{40} \cdot 100 = 30$$

Tabla 2-13 *Niveles ordenados de cotinina de 40 fumadores*

0	1	1	3	17	32	35	44	48	86
87	103	112	121	123	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	265	266	277	284	289	290	313	477	491

Notación

n = número total de valores en el conjunto de datos
k = percentil utilizado (ejemplo: para el percentil 25o, $k = 25$).
L = localizador que da la *posición* de un valor (ejemplo: para el valor 12o en la lista ordenada, $L = 12$).
P_k = percentil k-ésimo (ejemplo: P_{25} es el percentil 25o).

Desviación media absoluta: es la media de los valores absolutos de las desviaciones desde la media:

$$\textit{desviación media absoluta} = \frac{\textit{(suma de (valores absolutos de desviaciones)}}{\textit{número}}$$

$$\textit{desviación media absoluta} = \frac{\sum |x - \bar{x}|}{x} \qquad (2.5)$$

Varianza muestral: la varianza muestral, s^2, es la media del cuadrado de las desviaciones, calculada usando $n - 1$ como divisor:

$$\textit{varianza muestral: s cuadrada} = \frac{\textit{suma de (cuadrado de desviaciones)}}{\textit{número} - 1}$$

$$\mathbf{s^2} = \frac{\sum (x - x)^2}{n - 1} \qquad (2.6)$$

donde n es el tamaño muestral, es decir, el número de datos de la muestra.